D48–D53 *Nucleic Acids Research, 2012, Vol. 40, Database issue* *Published online 5 December 2011*
doi:10.1093/nar/gkr1202

GenBank

Dennis A. Benson, Ilene Karsch-Mizrachi, Karen Clark, David J. Lipman, James Ostell and Eric W. Sayers*

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Building 38A, 8600 Rockville Pike, Bethesda, MD 20894, USA

Received September 30, 2011; Revised November 14, 2011; Accepted November 17, 2011

ABSTRACT

GenBank® is a comprehensive database that contains publicly available nucleotide sequences for more than 250 000 formally described species. These sequences are obtained primarily through submissions from individual laboratories and batch submissions from large-scale sequencing projects, including whole-genome shotgun (WGS) and environmental sampling projects. Most submissions are made using the web-based BankIt or standalone Sequin programs, and accession numbers are assigned by GenBank staff upon receipt. Daily data exchange with the European Nucleotide Archive (ENA) and the DNA Data Bank of Japan (DDBJ) ensures worldwide coverage. GenBank is accessible through the NCBI Entrez retrieval system, which integrates data from the major DNA and protein sequence databases along with taxonomy, genome, mapping, protein structure and domain information, and the biomedical journal literature via PubMed. BLAST provides sequence similarity searches of GenBank and other sequence databases. Complete bimonthly releases and daily updates of the GenBank database are available by FTP. To access GenBank and its related retrieval and analysis services, begin at the NCBI home page: www.ncbi.nlm.nih.gov.

INTRODUCTION

GenBank (1) is a comprehensive public database of nucleotide sequences and supporting bibliographic and biological annotation. GenBank is built and distributed by the National Center for Biotechnology Information (NCBI), a division of the National Library of Medicine (NLM), located on the campus of the US National Institutes of Health (NIH) in Bethesda, MD, USA.

NCBI builds GenBank primarily from the submission of sequence data from authors and from the bulk submission of expressed sequence tag (EST), genome survey sequence (GSS), whole-genome shotgun (WGS) and other high-throughput data from sequencing centers. The US Office of Patents and Trademarks also contributes sequences from issued patents. GenBank participates with the EMBL Nucleotide Sequence Database (EMBL-Bank), part of the European Nucleotide Archive (ENA) (2), and the DNA Data Bank of Japan (DDBJ) (3) as a partner in the International Nucleotide Sequence Database Collaboration (INSDC). The INSDC partners exchange data daily to ensure that a uniform and comprehensive collection of sequence information is available worldwide. NCBI makes the GenBank data available at no cost over the Internet, through FTP and a wide range of web-based retrieval and analysis services (4).

RECENT DEVELOPMENTS

PopSet redesign

In the past year, NCBI redesigned the web interface for the PopSet database (www.ncbi.nlm.nih.gov/popset) of related sequences and alignments derived from phylogenetic, population, mutation and ecosystem studies that have been submitted to GenBank. The new PopSet record views contain three sections: an introduction showing the title and citation for the citation reporting the data set; a list of sequences contained in the data set; and, when available, an alignment of the sequences shown in the same Graphical Sequence Viewer that is a display option on nucleotide and protein records. In addition, PopSet record pages now display links to other PopSet records reported in the same published study, making it much easier to locate these related records. For PopSet records with fewer than 100 sequences, links are provided to generate a BLAST alignment of the sequences or, if an alignment was submitted as part of the record, a distance tree view of the alignment.

New tools for exploring features in GenBank records

Within any NCBI nucleotide or protein sequence record, the hyperlinks in the feature table section now open a new tool that highlights that feature in the sequence and displays details of that annotation. A bar will appear at

*To whom correspondence should be addressed. Tel: +301 496 2475; Fax: +301 480 9241; Email: sayers@ncbi.nlm.nih.gov

the bottom of the browser window that allows users to navigate to other features within the record or to view the sequence corresponding to the feature in FASTA or GenBank format. For example, the various segments of a CDS on a genomic sequence can be highlighted together and then displayed and downloaded as a single FASTA record. Sequence records also now have a 'Find in this Sequence' link in the right column that opens a search bar at the bottom of the browser window. Users can then input subsequences, including standard ambiguity codes and nucleotides and Prosite patterns for proteins, and then locate these patterns within the current sequence.

Unverified sequences

As part of the standard submission process, GenBank staff review submissions for biological accuracy and assist authors in providing accurate annotations. If GenBank staff is unable to verify the accuracy of the submitted sequences and/or annotations, they may now add a comment to the record stating that the sequence is unverified. Until the submitter is able to resolve the issues, such sequences will have the word 'UNVERIFIED:' at the beginning of their definition lines and will not be included in BLAST databases.

WGS browser

Within GenBank, WGS master records (see below) contain no sequence data, but rather show the descriptive information and range of accession numbers of the contigs submitted as part of that WGS project. In the future, NCBI will no longer assign GI numbers to these individual contigs, and at that point users will be unable to view them directly in the Nucleotide database. Instead, users may view these records in a WGS browser that is linked from the WGS feature of any WGS master record. The WGS browser provides the complete descriptive information from the master record of the project, interactive views of the FASTA of every contig record and also provides links to the FTP files for all the contigs of the entire project.

ORGANIZATION OF THE DATABASE

GenBank divisions

GenBank groups sequence records into various divisions based either on the source taxonomy or the sequencing strategy used to obtain the data. There are 12 taxonomic divisions (BCT, ENV, INV, MAM, PHG, PLN, PRI, ROD, SYN, UNA, VRL, VRT) and six high-throughput divisions (EST, GSS, HTC, HTG, STS, TSA). Finally, the PAT division contains records supplied by patent offices and the WGS division contains sequences from WGS projects. The size and growth of these divisions, and of GenBank as a whole, are shown in Table 1.

Sequence-based taxonomy

Database sequences are classified and can be queried using a comprehensive sequence-based taxonomy (www.ncbi.nlm.nih.gov/taxonomy/) developed by NCBI in collaboration with EMBL-Bank and DDBJ and with the valuable assistance of external advisers and curators. Almost 250 000 formally described species are represented in GenBank, and the top species in the non-WGS GenBank divisions are listed in Table 2.

Table 1. Growth of GenBank divisions (nucleotide base pairs)

Division	Description	Release 185 (8/2011)	Annual increase (%)[a]
TSA	Transcriptome shotgun data	1 874 047 448	370.1
ENV	Environmental samples	2 553 693 157	48.2
PHG	Phages	62 579 756	44.0
PAT	Patented sequences	11 154 487 762	30.9
BCT	Bacteria	6 975 597 755	30.8
INV	Invertebrates	2 535 336 197	24.5
WGS	Whole-genome shotgun data	208 315 831 132	23.1
VRL	Viruses	1 180 083 600	21.6
MAM	Other mammals	807 098 397	18.8
PLN	Plants	4 741 991 057	17.4
GSS	Genome survey sequences	20 770 772 329	12.6
SYN	Synthetic	156 218 063	9.6
VRT	Other vertebrates	2 705 250 711	6.8
EST	Expressed sequence tags	39 018 185 344	6.0
UNA	Unannotated	125 912	4.7
PRI	Primates	6 116 546 725	2.9
ROD	Rodents	4 396 957 541	2.3
HTC	High-throughput cDNA	662 320 919	0.4
STS	Sequence tagged sites	635 872 683	0.3
HTG	High-throughput genomic	24 324 068 445	0.2
TOTAL	All GenBank sequences	338 987 064 933	18.2

[a]Measured relative to Release 179 (8/2010).

Table 2. Top Organisms in GenBank (Release 185)

Organism	Non-WGS base pairs
Homo sapiens	15 881 839 899
Mus musculus	9 118 049 806
Rattus norvegicus	6 503 434 302
Bos taurus	5 381 235 474
Zea mays	5 055 840 446
Sus scrofa	4 793 300 236
Danio rerio	3 127 958 433
Strongylocentrotus purpuratus	1 352 948 327
Oryza sativa Japonica Group	1 251 053 810
Nicotiana tabacum	1 194 842 997
Xenopus (Silurana) tropicalis	1 147 237 486
Arabidopsis thaliana	1 138 511 865
Drosophila melanogaster	1 058 563 193
Pan troglodytes	1 003 309 475
Canis lupus familiaris	947 332 578
Vitis vinifera	915 431 680
Gallus gallus	896 784 038
Glycine max	895 052 594
Macaca mulatta	828 906 407
Solanum lycopersicum	778 132 243

Sequence identifiers and accession numbers

Each GenBank record, consisting of both a sequence and its annotations, is assigned a unique identifier called an accession number that is shared across the three collaborating databases (GenBank, DDBJ, EMBL Bank).

The accession number appears on the ACCESSION line of a GenBank record and remains constant over the lifetime of the record, even when there is a change to the sequence or annotation. Changes to the sequence data itself are tracked by an integer extension of the accession number, and this *Accession.version* identifier appears on the VERSION line of the GenBank flat file. The initial version of a sequence has the extension '.1'. In addition, each version of the DNA sequence is also assigned a unique NCBI identifier called a *GI* number that also appears on the VERSION line following the *Accession.version*:

ACCESSION AF000001
VERSION AF000001.5 GI : 7274584

When a change is made to a sequence in a GenBank record, a new GI number is issued to the updated sequence and the version extension of the *Accession.version* identifier is incremented. The accession number for the record as a whole remains unchanged and will always retrieve the most recent version of the record; the older versions remain available under the old *Accession.version* identifiers and their original GI numbers.

A similar system tracks changes in the corresponding protein translations. These identifiers appear as qualifiers for CDS features in the FEATURES portion of a GenBank entry, e.g. /protein_id = 'AAF14809.1'. Protein sequence translations also receive their own unique GI number, which appears as a second qualifier on the CDS feature:

/db_xref = 'GI : 6513858'

BUILDING THE DATABASE

The data in GenBank and the collaborating databases, EMBL-Bank and DDBJ, are submitted either by individual authors to one of the three databases or by sequencing centers as batches of EST, STS, GSS, HTC, WGS or HTG sequences. Data are exchanged daily with DDBJ and EMBL-Bank so that the daily updates from NCBI servers incorporate the most recently available sequence data from all sources.

Direct electronic submission

Virtually all records enter GenBank as direct electronic submissions (www.ncbi.nlm.nih.gov/genbank/), with the majority of authors using the BankIt or Sequin programs. Many journals require authors with sequence data to submit the data to a public sequence database as a condition of publication. GenBank staff can usually assign an accession number to a sequence submission within two working days of receipt, and do so at a rate of ~3500/day. The accession number serves as confirmation that the sequence has been submitted and provides a means for readers of articles in which the sequence is cited to retrieve the data. Direct submissions receive a quality assurance review that includes checks for vector contamination, proper translation of coding regions, correct taxonomy and correct bibliographic citations. A draft of the GenBank record is passed back to the author for review before it enters the database.

Authors may ask that their sequences be kept confidential until the time of publication. Since GenBank policy requires that the deposited sequence data be made public when the sequence or accession number is published, authors are instructed to inform GenBank staff of the publication date of the article in which the sequence is cited in order to ensure a timely release of the data. Although only the submitter is permitted to modify sequence data or annotations, all users are encouraged to report lags in releasing data or possible errors or omissions to GenBank at update@ncbi.nlm.nih.gov.

NCBI works closely with sequencing centers to ensure timely incorporation of bulk data into GenBank for public release. GenBank offers special batch procedures for large-scale sequencing groups to facilitate data submission, including the program *tbl2asn*, described at www.ncbi.nlm.nih.gov/genbank/tbl2asn2.html.

Submission using BankIt

About a third of author submissions are received through an NCBI web-based data submission tool named BankIt. Using BankIt, authors enter sequence information and biological annotations, such as coding regions or mRNA features, directly into a series of tabbed forms that allow the submitter to describe the sequence further without having to learn formatting rules or controlled vocabularies. Additionally, BankIt allows submitters to upload source and annotation using tab-delimited tables. Before creating a draft record in the GenBank flat file format for the submitter to review, BankIt validates the submissions by flagging many common errors and checking for vector contamination using a variant of BLAST called Vecscreen.

Submission using Sequin and tbl2asn

NCBI also offers a standalone multi-platform submission program called Sequin (www.ncbi.nlm.nih.gov/projects/Sequin/) that can be used interactively with other NCBI sequence retrieval and analysis tools. Sequin handles simple sequences (such as a single cDNA), phylogenetic studies, population studies, mutation studies, environmental samples with or without alignments and sequences with complex annotation. Sequin has a number of wizards that guide the submitter in preparing their submission with proper annotation for a number of data types, like viral genomic sequences and ribosomal RNA from cultured and uncultured microbes. Sequin has convenient editing and complex annotation capabilities and contains a number of built-in validation functions for quality assurance. In addition, Sequin is able to accommodate large sequences, such as the 5.6 Mb *Escherichia coli* genome, and read in a full complement of annotations from simple tables. The most recent version, Sequin 11.7, was released in September 2011 and is available for Macintosh, PC and Unix computers via anonymous FTP at ftp.ncbi.nih.gov/sequin. Once a submission is completed, submitters can e-mail the Sequin file to gb-sub@ncbi.nlm

.nih.gov. Submitters of large, heavily annotated genomes may find it convenient to use *tbl2asn* to convert a table of annotations generated from an annotation pipeline into an ASN.1 (Abstract Syntax Notation One) record suitable for submission to GenBank.

Submission of barcode sequences

The Consortium for the Barcode of Life (CBOL) is an international initiative to develop DNA barcoding as a tool for characterizing species of organisms using a short DNA sequence. For animal species, a 648-bp fragment of the gene for cytochrome oxidase subunit I is used as the barcode. The plant and fungal communities are using other loci. NCBI, in collaboration with CBOL (www.barcoding.si.edu/), provides an online tool (BarSTool) for the bulk submission of barcode sequences to GenBank (www.ncbi.nlm.nih.gov/WebSub/?tool=barcode) that allows users to upload files containing a batch of sequences with associated source information. The Nucleotide query 'barcode[keyword]' retrieves the over 500 000 barcode sequences in GenBank, over 300 000 of which were added in the last year.

Notes on particular divisions

Transcriptome shotgun assembly sequences. The TSA division contains transcriptome shotgun assembly (TSA) sequences that are assembled from sequences deposited in the NCBI Trace Archive, the Sequence Read Archive (SRA) and the EST division of GenBank. While neither the Trace Archive nor SRA is a part of GenBank, they are part of the INSDC and provide access to the data underlying these assemblies (4,5). TSA records (for example, EZ000001) have 'TSA' as their keyword and can be retrieved with the query 'tsa[properties]'.

Environmental sample sequences. The environmental sample sequences (ENV) division of GenBank accommodates sequences obtained via environmental sampling methods in which the source organism is unknown. Many ENV sequences arise from metagenome samples derived from microbiota in various animal tissues, such as within the gut or skin, or from particular environments, such as freshwater sediment, hot springs or areas of mine drainage. Records in the ENV division contain 'ENV' in the keyword field and use an '/environmental_sample' qualifier in the source feature.

WGS sequences. WGS sequences appear in GenBank as groups of sequence-overlap contigs collected under a master WGS record. Each master record represents a WGS project and has an accession number in the Nucleotide database consisting of a four-letter prefix followed by eight zeroes and a version suffix as found in standard GenBank records. The number of zeroes increases to nine for WGS projects with one million or more contigs. Master records contain no sequence data; rather, they are linked to their set of individual contigs that can be viewed using the new WGS browser (see above). Contig records have accessions consisting of the same four-letter prefix as their master accession, followed by a two-digit version number and a six-digit contig ID. For example, the WGS accession number 'AAAA02002744' is assigned to contig number '002744' of the second version of project 'AAAA', whose accession number is 'AAAA00000000.2'. Currently, there are over 3400 WGS sequencing projects, many of whose data have been used to build more than 9 million scaffolds and chromosomes for genome assemblies. For a complete list of WGS projects with links to the data, see www.ncbi.nlm.nih.gov/Traces/wgs/.

Although WGS project sequences may be annotated, many low-coverage genome projects do not contain annotation. Because these sequence projects are ongoing and incomplete, these annotations may not be tracked from one assembly version to the next and should be considered preliminary. Submitters of genomic sequences, including WGS sequences, are urged to use evidence tags of the form '/experimental = *text*' and '/inference = *TYPE*:*text*', where *TYPE* is one of a number of standard inference types and *text* consists of structured text.

Expressed sequence tags. Expressed sequence tags (ESTs) continue to be a major source of data for gene expression and annotation studies, and at 39 billion base pairs, it remains the largest non-WGS division in GenBank. EST data are available for download from ftp.ncbi.nih.gov/repository/dbEST/ (6) as well as from the GenBank FTP site. The data in dbEST are clustered using the BLAST programs to produce the UniGene database (www.ncbi.nlm.nih.gov/unigene) of more than 5.3 million gene-oriented sequence clusters representing almost 140 organisms (4).

High-throughput genomic and high-throughput cDNA sequences. The high-throughput genomic (HTG) division of GenBank (www.ncbi.nlm.nih.gov/genbank/htgs/) contains unfinished large-scale genomic records, which are in transition to a finished state (7). These records are designated as belonging to Phases 0–3 depending on the quality of the data, with Phase 3 being the finished state. Upon reaching Phase 3, HTG records are moved into the appropriate organism division of GenBank.

The HTC division of GenBank contains high-throughput cDNA (HTC) sequences that are of draft quality but may contain 5′-UTRs, 3′-UTRs, partial coding regions and introns. HTC sequences which are finished and of high quality are moved to the appropriate organism division of GenBank. A project generating HTC data is described in Ref. (8).

Special record types

Third party annotation. Third party annotation (TPA) records are sequence annotations published by someone other than the original submitter of the primary sequence record in DDBJ/EMBL/GenBank (www.ncbi.nlm.nih.gov/genbank/TPA). Each of the current 160 000 TPA records falls into one of three categories: *experimental*, in which case there is direct experimental evidence for the existence of the annotated molecule; *inferential*, in which case the experimental evidence is indirect; and

reassembly, where the focus is on providing a better assembly of the raw reads. TPA sequences may be created by assembling a number of primary sequences. The format of a TPA record (e.g. BK000016) is similar to that of a conventional GenBank record but includes the label 'TPA_exp:', 'TPA_inf:' or 'TPA_reasm:' at the beginning of each Definition Line as well as corresponding keywords. TPA experimental and inferential records also contain a Primary block that provides the base ranges and identifier for the sequences used to build the TPA. TPA sequences are not released to the public until their accession numbers or sequence data and annotation appear in a peer-reviewed biological journal. TPA submissions to GenBank may be made using either BankIt or Sequin.

Contig (CON) records for assemblies of smaller records. Within GenBank, CON records are used to represent very long sequences, such as a eukaryotic chromosome, where the sequence is not complete but consists of several contig records with uncharacterized gaps between them. Rather than listing the sequence itself, CON records contain assembly instructions involving the several component sequences. An example of such a CON record is CM000663 for human chromosome 1.

RETRIEVING GENBANK DATA

The Entrez system

The sequence records in GenBank are accessible through the NCBI Entrez retrieval system (4). Records from the EST and GSS divisions of GenBank are stored in the EST and GSS databases, while all other GenBank records are stored in the Nucleotide database. GenBank sequences that are part population or phylogenetic studies are collected together in the PopSet database, and conceptual translations of CDS sequences annotated on GenBank records are available in the Protein database. Each of these databases is linked to the scientific literature via PubMed and PubMed Central. Additional information about conducting Entrez searches is found in the NCBI Help Manual (www.ncbi.nlm.nih.gov/books/NBK3831/) and links to related tutorials are provided on the NCBI Education page (www.ncbi.nlm.nih.gov/Education/).

Associating sequence records with sequencing projects

The ability to identify all GenBank records submitted by a specific group or those with a particular focus, such as metagenomic surveys, is essential for the analysis of large volumes of sequence data. The use of organism or submitter names as a means to define such a set of sequences is unreliable. The BioProject database (www.ncbi.nlm.nih.gov/bioproject), developed at NCBI and subsequently adopted across the INSDC, allows submitters to register large-scale sequencing projects under a unique project identifier, enabling reliable linkage between sequencing projects and the data they produce. BioProject, which replaced the Genome Projects database, is broader in scope than its predecessor and includes pointers to data from a wide variety of projects deposited in any NCBI primary data archive. Sequencing projects focus on genomes, metagenomes, transcriptomes, comparative genomics as well as on particular loci, such as 16S ribosomal RNA. A 'DBLINK' line appearing in GenBank flat files identifies the sequencing projects with which a GenBank sequence record is associated. As an example, the DBLINK line below associates a GenBank sequence record with Project record 18787.

DBLINK Project:18787

In the future these links will change to reflect the new BioProject accessions: 'DBLINK BioProject:PRJNA 18787'. Project record 18787 provides details of the sequencing progress for the green anole, *Anolis carolinensis* (www.broad.mit.edu/models/anole/). Within the Entrez system, such a sequence record is linked directly to the appropriate BioProject record; these links are bidirectional, so that the BioProject records also link back to associated sequence records.

BLAST sequence-similarity searching

Sequence-similarity searches are the most fundamental and frequent type of analysis performed on GenBank data. NCBI offers the BLAST family of programs (blast.ncbi.nlm.nih.gov) to detect similarities between a query sequence and database sequences (9,10). BLAST searches may be performed on the NCBI website (11) or by using a set of standalone programs distributed by FTP (4).

Obtaining GenBank by FTP

NCBI distributes GenBank releases in the traditional flat file format as well as in the ASN.1 format used for internal maintenance. The full bimonthly GenBank release along with the daily updates, which incorporate sequence data from EMBL-Bank and DDBJ, is available by anonymous FTP from NCBI at ftp.ncbi.nih.gov/genbank. The full release in flat file format is available as a set of compressed files with a non-cumulative set of updates at http://ftp.ncbi.nih.gov/genbank/daily-nc/. For convenience in file transfer, the data are partitioned into multiple files; for release 185 there are 1659 files requiring 550 GB of uncompressed disk storage. A script is provided in ftp.ncbi.nih.gov/genbank/tools/ to convert a set of daily updates into a cumulative update.

ELECTRONIC ADDRESSES

www.ncbi.nlm.nih.gov: NCBI Home Page.
gb-sub@ncbi.nlm.nih.gov: Submission of sequence data to GenBank.
update@ncbi.nlm.nih.gov: Revisions to, or notification of release of, 'confidential' GenBank entries.
info@ncbi.nlm.nih.gov: General information about NCBI resources.

CITING GENBANK

If you use the GenBank database in your published research, we ask that this article be cited.

FUNDING

Funding for open access charge: Intramural Research Program of the National Institutes of Health, National Library of Medicine.

Conflict of interest statement. None declared.

REFERENCES

1. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Sayers,E.W. (2010) GenBank. *Nucleic Acids Res.*, **38**, D46–D51.
2. Leinonen,R., Akhtar,R., Birney,E., Bower,L., Cerdeno-Tarraga,A., Cheng,Y., Cleland,I., Faruque,N., Goodgame,N., Gibson,R. *et al.* (2011) The European Nucleotide Archive. *Nucleic Acids Res.*, **39**, D28–D31.
3. Kaminuma,E., Kosuge,T., Kodama,Y., Aono,H., Mashima,J., Gojobori,T., Sugawara,H., Ogasawara,O., Takagi,T., Okubo,K. *et al.* (2011) DDBJ progress report. *Nucleic Acids Res.*, **39**, D22–D27.
4. Sayers,E.W., Barrett,T., Benson,D.A., Bolton,E., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., Dicuccio,M., Federhen,S. *et al.* (2012) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **40**, D13–D25.
5. Shumway,M., Cochrane,G. and Sugawara,H. (2010) Archiving next generation sequencing data. *Nucleic Acids Res.*, **38**, D870–D871.
6. Boguski,M.S., Lowe,T.M. and Tolstoshev,C.M. (1993) dbEST—database for "expressed sequence tags". *Nat. Genet.*, **4**, 332–333.
7. Kans,J.A. and Ouellette,B.F.F. (2001) Submitting DNA Sequences to the Databases. In: Baxevanis,A.D. and Ouellette,B.F.F. (eds), *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*. John Wiley and Sons, Inc., New York, NY, pp. 65–81.
8. Kawai,J., Shinagawa,A., Shibata,K., Yoshino,M., Itoh,M., Ishii,Y., Arakawa,T., Hara,A., Fukunishi,Y., Konno,H. *et al.* (2001) Functional annotation of a full-length mouse cDNA collection. *Nature*, **409**, 685–690.
9. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
10. Zhang,Z., Schaffer,A.A., Miller,W., Madden,T.L., Lipman,D.J., Koonin,E.V. and Altschul,S.F. (1998) Protein sequence similarity searches using patterns as seeds. *Nucleic Acids Res.*, **26**, 3986–3990.
11. Johnson,M., Zaretskaya,I., Raytselis,Y., Merezhuk,Y., McGinnis,S. and Madden,T.L. (2008) NCBI BLAST: a better web interface. *Nucleic Acids Res.*, **36**, W5–W9.